The 65nm 16MB On-die L3 Cache for a Dual Core Multi-Threaded Xeon® Processor

Jonathan Chang, Ming Huang, Jonathan Shoemaker, John Benoit, Szu-Liang Chen, Wei Chen, Siufu Chiu, Raghuraman Ganesan, Gloria Leong, Venkata Lukka, Stefan Rusu, Durgesh Srivastava

Intel Corporation
2200 Mission College Blvd. (SC12-408), Santa Clara, CA 95052, USA.

ABSTRACT

The 16-way set associative, single-ported 16MB cache for the dual-core Xeon® Processor uses a 0.624um^2 cell in a 65nm 8-metal technology. Only 0.8% of the cache is powered up for an access. Sleep transistors are used in the SRAM array and peripherals. Dynamic Pellston with a history buffer protects the cache from latent defects and infant mortality failures.

INTRODUCTION

The dual-core Xeon® Processor with a 16MB unified L3 cache is implemented in a 65nm process technology with 8 copper interconnect layers [1]. It consists of two cores, each with a 1MB L2 cache. Both L3 and L2 use the same 0.624 μm^2 bit cell. Sleep transistors were designed in SRAM arrays and their peripherals to achieve 0.75W/MB average power, while maintaining the cache content all the time [2]. Shutdown option is implemented in the SRAM arrays to minimize the leakage power for the inactive sub arrays. Pellston technology is enhanced to keep track of the random ECC event of each cache line and disable the cache lines susceptible to latent defects and infant mortality [3].

CACHE ORGANIZATION

The L3 cache is a 16-set, 16 way set-associative cache, organized as shown in Fig. 1.

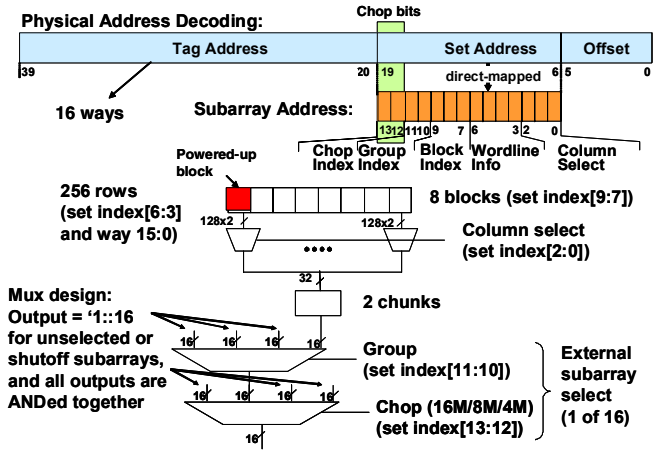

Figure 1-Cache Organization

For each cache access, only one of the 16 sub arrays in each group is accessed, and one of the 8 blocks in the accessed sub array is powered up. As a result, only 0.8% of all array blocks power up for each cache access. Inactive sub arrays drive 1's to their local data buses. AND structure is implemented as the multiplexer to minimize the gate delay. The cache line size is 64B, which is sent in 2 chunks on the data buses. Each chunk has 256 data bits, 32 ECC bits and 2 redundancy bits. The L3 cache uses 256 64KB regular sub arrays and 32 68KB redundancy sub arrays. Each regular sub array stores 32 bits of data, while the redundancy sub array stores 34 bits. Each 16 sub arrays form one group, with 18 data groups overall. Fig. 2 shows the construction of a 64K/68K sub array and the power-up resolution. Each sub array has 8 blocks, and the power up resolution is designed at the block level.

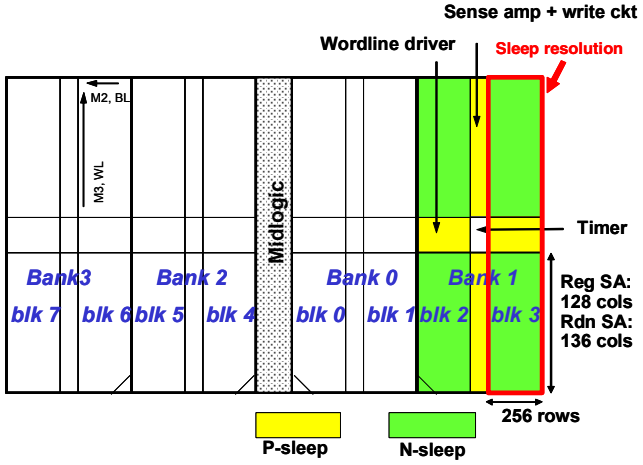

Figure 2-64KB/68KB Sub array

Cache sizing can be accomplished through both set reduction and way reduction. Set reduction is available when the cache size is cut down to either 8MB or 4MB. Way-reduction can be used to achieve intermediate cache sizes. Table 1 shows the cache organization for 16MB, 8MB, and 4MB options.

Table 1-Cache Sizing Option

Cache Size	16 MB	8 MB	4 MB
Set	16 K	8 K	4 K
Index	(14) PA[19:6]	(13) PA[18:6]	(12) PA[17:6]
Tag	(20) PA[39:20]	(21) PA[39:19]	(22) PA[39:18]
Way	16	16	16

LEAKAGE POWER REDUCTION

The SRAM arrays have a NMOS sleep function as shown in Fig. 3. A biasing circuitry controls the temperature variation of the virtual ground. The sleep bias is programmable and it was characterized such that the variation of the virtual ground across process corners and temperature is within specification. An NMOS diode and a PMOS pulldown were added to control the temperature variation of the virtual ground. When virtual VSS goes above transistor threshold voltage (Vt), MND will be on because Vgs is larger than Vt. MPB pulldown function will also be more significant because Vds is larger than Vt. As a result, the virtual VSS will be limited to a Vt or less. During shutoff, the sleep bias is cut off from the ground and the PMOS diode is turned off to let the virtual ground float up to the maximum value.

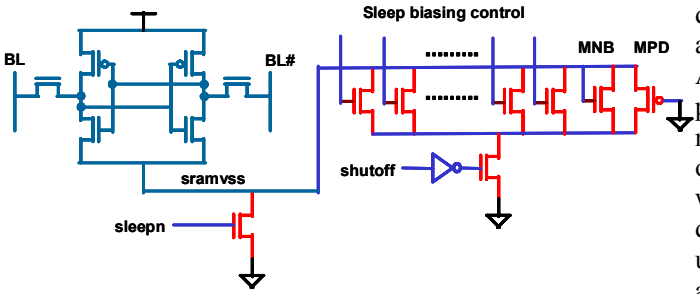

Figure 3-Sleep Biasing Circuitry

PMOS sleep is implemented in the decoders and cache i/o, which include column muxes, write drivers, and sense amplifiers. Fig. 4 shows the p-sleep design in the decoder and WL drivers. Fig. 5 shows the p-sleep design in the cache i/o.

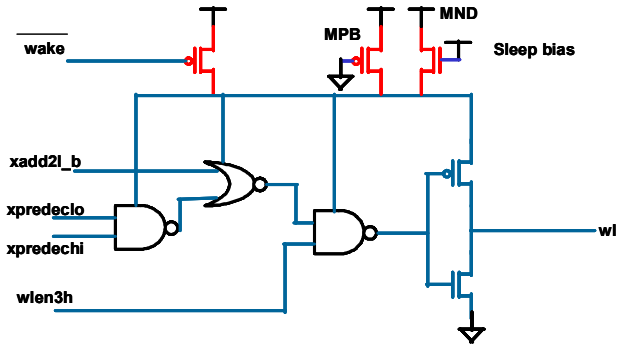

Figure 4-Decoder and WL driver with p-sleep design

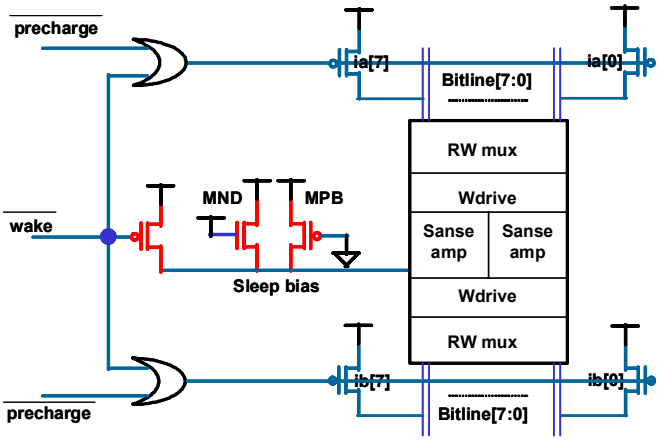

Figure 5-Cache I/O with p-sleep design

The bitlines are floated when the cache i/o is in sleep mode to reduce the bitline leakage. This also avoids DC current across the read/write column mux when they are driven by virtual Vcc. The virtual Vcc of the decoders and cache i/o is a fixed setting and is limited to be within a predetermined drop from nominal Vcc so that proper logic values are available at the wordlines during the sleep mode. An NMOS diode is implemented to limit the voltage drop for virtual Vcc. The bitline prechargers are excluded from the sleep transistor design to avoid the long precharge time and to meet the stringent requirements of bitline equalization for the differential sense amplifiers. The timer is intentionally left out of the sleep design because of its criticality of timing. Long Le transistors are used wherever possible in the timer without compromising the accuracy of the signal edges. With the p-sleep design in the cache peripherals, the incremental power saving is approximately 6W.

All 16 ways are included in a block, primarily to hide the latency penalty of turning on sleep transistors and to minimize the number of blocks/sub arrays to be powered up. Each block has one wake signal, which is used for both the SRAM sleep and wordline sleep. To minimize the dynamic power consumption due to the switching of the sleep transistors, programmable wake-up counters are implemented to detect if there is another cache access to the same sub array block within a pre-determined number of cycles. The default cycle count is chosen by running performance simulations for typical applications.

To reduce the dynamic power consumption, about 40% of the clock loading is gated in each sub array. The overall dynamic power consumption for the entire L3 cache is 1.7W for average applications. The switching power of n-sleep and p-sleep transistors is less than 100mW because of the low activity factor of L3 access and the small number of powered-up sub arrays for each access. Meanwhile, it takes about 49.7ns for virtual ground to reach the target value based on a pessimistic analysis. The efficiency of the leakage power saving is about 95%.

PELLSTON TECHNOLOGY

Dynamic Pellston was implemented to resolve Vccmin sensitivity from latent defects. An on-die history table (Pellston Engine Queue) keeps track of random ECC events for each cache line. Unlike the previous implementation [3], this dynamic Pellston can isolate random cache errors.

The first time an ECC error occurs on a cache line it could be a soft-error. The second occurrence of ECC error to the same location means that it is less likely to be a soft-error, but more likely a physical issue, such as a latent defect or Vccmin sensitivity. Pellston is enabled during both power-on self-test and in normal operation. Whenever one bit ECC error is detected, the ECC logic signals the error information to the Pellston Engine, including the set and way information for the Global Bus Queue (GBSQ). If the error is for the first time on the particular set and way, then entry is logged in the Pellston Engine Queue. Cache way is not invalidated or scrubbed on the first error. If the error is the second one on the same location, Pellston engine kicks in and disables the corresponding cache line.

ACKNOWLEDGMENTS

The authors would like to thank Kevin Zhang and his team for providing the technical feedback on p-sleep design and the baseline sub array design.

REFERENCES

[1] S. Rusu, et al., "A Dual Core Multi Threaded Xeon® Processor with 16MB L3 Cache," 2006 IEEE International Solid-State Circuits Conference.

[2] K. Zhang, et al., "A SRAM Design on 65nm CMOS Technology with Integrated Leakage Reduction Scheme," VLSI Symposium, 2004, pp 294–295.

[3] J. Wuu, et al., "The Asynchronous 24MB On-Chip Level-3 Cache for a Dual-Core Itanium® Family Processor," 2005 IEEE International Solid-State Circuits Conference, pp 488-489.